学校编码：10384　　分类号_____密级 _____
学　　号：200228051　　UDC_____

厦　门　大　学

硕　士　学　位　论　文

对齐模板统计翻译模型中的双语词聚类研究

The Research of Bilingual Word Clustering in Alignment Template Statistical Translation Model

张 振 昌

指导教师姓名：　史　晓　东　教授
专　业　名　称：　计　算　机　应　用
论文提交日期：　2005　年　　月
论文答辩日期：　2005　年　　月
学位授予日期：　2005　年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2005 年　　月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权利和责任。

声明人（签名）：

年　　月　　日

摘　要

基于统计的方法是当前机器翻译领域主流的研究方向之一，其中对齐模板统计翻译模型是效果最好的一个统计模型，而双语词聚类在对齐模板统计模型参数训练中占有十分重要的地位。

本文首先综述了目前统计机器翻译的研究现状，简要介绍了几种不同的统计机器翻译模型。接着详细介绍了对齐模板的统计翻译模型及其模型参数训练。

针对参数训练中的双语词聚类，在原来非层次聚类的基础上，给出了双语层次聚类的算法，同非层次聚类相比聚类的效果有一定的提高；又结合了两种算法的思想，提出了一种新的混合算法，新算法聚类的效果又有进一步的提高。为了确定最优的聚类个数，引入了最小描述长度的评价标准，对不同聚类个数的多个模型，选择其中描述长度最小的模型为最优模型，该模型的聚类个数就是最优的聚类个数。尝试使用 Fuzzy K-means 算法对双语单词进行模糊聚类，针对自然语言的特点，探讨了单词(对象)属性向量的选择以及距离函数的选择。

最后我们完成对齐模板其他参数的训练，把参数训练结果和解码器整合在一起，实现了对齐模板的统计翻译系统。

从实验结果中，相对于层次和非层次聚类，我们用合并算法得到的聚类结果，应用于对齐模板的统计翻译系统，系统的翻译性能有一定的提高。

关键词：统计机器翻译，对齐模板，双语词聚类，最小描述长度，模糊 K-means

Abstract

Statistical machine translation is one of the main research directions in machine translation research field. The statistical model of alignment template is the one of best of statistical machine translation models. And bilingual word clustering plays a key role in the parameter training of alignment template.

Firstly, this thesis summarized the status of statistical machine translation and introduced briefly some statistical machine translation models. Then the alignment template statistical model and the parameter training of this model were presented in detail.

Secondly, for training the bilingual word clustering, we presented bilingual hierarchical clustering algorithm. Compared with bilingual flat clustering algorithm, the result of hierarchical clustering was better. By combining two algorithms, we presented a new merge algorithm and resulted in a further gain. In order to determine the optimal number of clusters, Minimum Description Length was regarded as evaluation standard. Fuzzy K-means algorithm was employed in the processing of clustering on bilingual corpus. Then we discussed the feature vector of a word (object) and the distance function between two objects.

Finally, we accomplished the training of other parameters training of alignment template. Combining the training result and the decoder, we achieved an alignment template statistical machine translation system.

The result showed: the performance of the machine translation system was slightly improved, after applying the clustering result trained by merge algorithm to the system.

Keywords: statistical machine translation; alignment template; bilingual word clustering; minimum description length; fuzzy k-means.

目　录

第一章　引言

1.1 统计机器翻译概述[1]

上世纪九十年代初，IBM 的 Peter Brown 等人提出基于信源信道思想的统计机器翻译模型[2]，而且在实验上取得初步的成功。不过由于当时计算能力不足等多方面，对统计机器翻译方法进行深入研究的人并不多，而且 IBM 在 1996 年又忽然放弃了在这方面的研究，人们对统计机器翻译的有效性还持有怀疑的态度。直到 1999 年，约翰霍金斯(JHU)大学的统计机器翻译夏季讨论班[3]汇聚了许多研究者，大家重复了 IBM 当年的实验，并开发出了一套开放源代码的统计机器翻译工具包——Egypt[4]。在这之后研究者回到各自的研究机构，对 IBM 的统计机器翻译模型提出各种的改进，这才掀起了统计机器翻译新一轮的研究热潮。到了现在，统计方法如日中天，已经成为国际上机器翻译的主流方法之一。

IBM 的统计机器翻译方法的基本思想是，把机器翻译看成一个信息传输的过程，用一种信源信道模型对机器翻译进行解释。假设一个目标语言句子 e，经过某一噪声信道后变成了源语言句子 f，也就是假设源语言句子 f 通过某种编码后得到了目标语言句子 e，而我们翻译的目的就是要将 f 还原成 e。

利用贝叶斯公式，可以得到下面的公式

$$\Pr(e \mid f) = \frac{\Pr(f \mid e) \cdot \Pr(e)}{\Pr(f)}$$

翻译的目标就是对给定的源语言句子f找到对应最好的目标语言句子$\hat{e}$，而分母部分$\Pr(f)$和目标语言e无关，因此只要最大化分子部分$\Pr(f|e)\Pr(e)$：

$$\hat{e} = \arg\max_{e} \Pr(f \mid e) \cdot \Pr(e)$$

这个公式被Brown等人称为**统计机器翻译的基本公式**[2]。

在这个公式中Pr(*e*)，是目标语言的句子出现的概率，称为语言模型。Pr(*f*|*e*)是由目标语言句子*e*翻译成源语言句子*f*的概率，称为翻译模型。语言模型只与目标语言相关，与源语言无关，体现的是一个目标句子在目标语言中出现的可能性，也就是该目标语言句子在句法语义等方面的合理程度，反映了译文句子的流利度；翻译模型与源语言和目标语言都有关系，体现了两个句子互译的可能性，反映了忠实度。

从上面的公式，统计机器翻译可以分解成三个问题：

1) 语言模型的参数训练

2) 翻译模型的参数训练

3) 搜索问题，有效快速地找到最优或者次优的译文

在估算译文的时候，为什么不直接使用Pr(*e*| *f*)，而使用Pr(*f*| *e*)Pr(*e*)这样一个更复杂的公式呢？

如果直接使用Pr(*e*| *f*)来选择目标句子，能够把上面的三个问题减少成两个，但是得到的结果*e*很有可能是不符合目标语言的语法的；采用了Pr(*f*| *e*)Pr(*e*)作为基本公式，语言模型Pr(*e*)就可以保证得到的译文尽可能的符合目标语言语法。

1.1.1 语言模型

对于语言模型，一般采用n元模型，也可以使用链语法等语法模型。链语法模型相比n元模型的优点在于可以处理长距离的依赖关系。

1.1.2 翻译模型

Brown等人提出了五个翻译模型，分别称为模型1-5[5]。模型1中只考虑单词之间的对译概率$p(f_j|e_{aj})$，单词的位置对模型没有影响，j是源语言词的

位置，a_j是与f_j对齐的目标语言词e_{aj}的位置。模型2 中，考虑了翻译过程单词的位置变化，引入参数$p(a_j \mid j, J, I)$，I和J分别是目标语言和源语言句子的长度。模型3考虑了一个单词翻译成多个单词的情况，引入了产生概率(*fertility*) $p(\phi_i \mid e_i)$，表示单词e_i翻译成ϕ_i个源语言单词的概率。模型4在对齐时不仅仅考虑词的位置变化，同时考虑了该位置上单词的词类(基于词类的对齐模型，把源语言和目标语言的单词自动划分到50个词类中)。模型5是对模型4的修正，消除了模型4中的缺陷，避免了对不可能出现对齐给出非零的概率。

在模型1和模型2中，假设所有长度都具有相同的可能性；对于源语言句子中的每个位置，猜测其与目标语言单词的对应关系，以及该位置上的源语言单词。在模型3、4、5中，首先，对于每个目标语言单词，根据$p(\phi_i \mid e_i)$选择与其对齐的源语言单词个数；然后再确定这些单词；最后，判断这些源语言单词的具体位置。

这些模型的主要区别在于计算源语言单词和目标语言单词之间的对齐概率计算方法不同。模型1最简单，只考虑词与词之间互译的概率，不考虑词的位置信息，也就是说，与词序无关。模型1的参数估计具有全局最优的特点，也就是说最后总可以收敛于一个与初始值无关的点；模型2也可以找到最优的对齐；模型3到5都只能收敛到局部最优。但在IBM的实验中，每一种模型的参数估计都是使用前一个模型的结果作为初始值，这样可以认为最后模型得到的结果也是和初始值无关的。

1.1.3 搜索问题

对与搜索问题，搜索空间一般是随着句子长度呈指数级别增长，是一个典型的NP-Hard问题，在多项式时间内找到全局最优解是不可能的。要在可以接受的时间限制内找到一个较好的译文，就要采用各种启发式搜索策略。

在IBM的实验中，Brown等人使用了一种称为堆栈搜索方法的变体[2]，这种方法在语音识别方面取得很好的效果。

IBM的试验表明，统计机器翻译的性能瓶颈并不在搜索问题。如果统计机器翻译的错误分为两种：

1) 模型错误：根据模型计算出来最高概率的译文并不是正确译文。
2) 搜索错误：虽然据模型计算出概率最高的译文是正确译文，但搜索算法由于使用各种的启发式策略，过早就错误地把正确译文排除在搜索范围之内，使得最终没有找到这个正确译文。

根据IBM的试验，后一类的错误只占所有翻译错误中的5%[6]。

IBM提出的统计机器翻译基本方程式对于机器翻译领域具有开创性的意义，而IBM的其他工作只是对这个基本方程式的一种诠释。从理论上讲，IBM的几个模型只考虑了词与词之间的线性关系，没有考虑句子的结构。对于语序差别比较大的语言对而言，效果可能会不大好。如果在考虑语言模型和翻译模型时将句法结构或语义结构考虑进来，应该会得到更好的结果。

1.2 Och 对齐模板的翻译模型[7]

在德国主持开发的著名的语音机器翻译系统Verbmobil中，Och所在的研究组承担了其中统计机器翻译模块。与IBM的模型相比，他们主要做了以下改进：

1）采用了基于类的模型。扩展单语聚类为双语聚类，利用非层次的聚类方法，通过多次迭代将两种语言的每一个词都对应到某个类中去。

2）在翻译模型上，采用了称为对齐模板的方法，实现了两种层次的对齐：短语层次的对齐和词语层次的对齐。对齐模板采用了基于类的对齐矩阵。对齐模板的获取是自动进行的，在对训练语料进行两个方向的IBM模型

词对齐以后，合并两个对齐，得到一个多对多的对齐矩阵，之后所有可能的对齐模板都被保存下来，并根据其在语料库种出现的频率赋予不同的概率。

3) 为了在搜索过程，更加容易地整合翻译模型和语言模型，简化搜索问题，他们对IBM提出的统计机器翻译的基本公式进行了修改，用反向翻译模型取代正常的翻译模型：

$$\hat{e} = \arg\max_{e} \Pr(e \mid f) \cdot \Pr(e)$$

实验的结果发现，这种改变并没有降低翻译的正确率[7]。

1.3 Koehn 的词组统计翻译系统[8]

Keohn实现了一个基于词组的统计翻译系统pharaoh，作为其博士毕业论文的一部分。与前面介绍的模型相比，该模型有以下的特点：

1) 在翻译模型中，基本的翻译对象不再是单词，而是词组。与Och对齐模板不同的是，pharaoh系统中不使用词类，在词组的每个位置上不是对应单词的词类，而是单词自身。pharaoh系统与对齐模板另外一个不同的地方是不存在两个层次的对齐，只有短语间的对齐，而没有短语内部的对齐。词组对的获取则与对齐模板类似，也是先进行两个方向的训练，然后合并对齐矩阵，最后抽取其中的短语对。

2) 为了校准输出译文的长度，Koehn引入了一个长度因子 w：

$$\hat{e} = \arg\max_{e} \Pr(e \mid f) \cdot \Pr(e) \cdot w^{length(e)}$$

1.4 本文的结构

本文的主要内容分为七章：

第一章是引言，首先介绍了统计机器翻译的基本思想，然后简要地综述IBM 模型 1-5 的翻译模型，接着又介绍了几个针对 IBM 改进的模型。

第二章介绍了研究的前提和基础。首先给出了 IBM 五个翻译模型的数学描述和模型参数的训练方法，接着介绍了 Och 的对齐模板的统计翻译模型，并给出其需要进行的参数训练。

第三章针对对齐模板模型参数训练中双语单词聚类，给出了双语单词聚类的目标函数，实现了 Och 提出双语聚类算法，在此基础上，又实现了双语的层次聚类算法，得到的聚类结果较原来的算法要好；又在两种算法之上，提出了一种新的混合算法，聚类的效果又有了进一步的提高。

第四章利用最小描述长度(MDL)的模型选择准则，对不同聚类个数的多个模型，选择其中描述长度最小的一个模型，作为最优的模型。

第五章我们尝试用 Fuzzy K-means 算法对双语单词进行模糊聚类，针对自然语言处理的特点，探讨了单词(对象)属性向量的表示方法以及距离函数的选择，并最终实现了该算法。

第六章首先介绍并完成了其他参数训练的结果，接着介绍了对齐模板翻译模型解码器的具体算法，最后整合参数训练结果和解码器，实现了对齐模板统计翻译系统。

第七章是结语，简单总结了目前所做的工作，并提出了进一步研究的设想。

1.5 小结

在这一章中，我们介绍了统计机器翻译的基本思想，并给出了统计机器翻译的基本公式，指出了统计机器翻译的三个基本问题：语言模型、翻译模型和搜索问题，又简要描述了IBM提出的五个翻译模型，这五个模型从简单模型1、模型2开始，可以快速地找到模型的最优对齐，逐渐变得复杂，